

A consistent deterministic regression tree for non-parametric prediction of time series

Pierre Gaillard^{1,2} and Paul Baudin³

¹ EDF R&D, Clamart, France

² GREGHEC (HEC Paris, CNRS), Jouy-en-Josas, France

`pierre-p.gaillard@edf.fr`

³ Inria, Roquencourt, France

`paul.baudin@inria.fr`

Abstract. We study online prediction of bounded stationary ergodic processes. To do so, we consider the setting of prediction of individual sequences and build a deterministic regression tree that performs asymptotically as well as the best L -Lipschitz constant predictors. Then, we show why the obtained regret bound entails the asymptotical optimality with respect to the class of bounded stationary ergodic processes.

1 Introduction

We suppose that at each time step $t = 1, 2, \dots$, the learner is asked to form a prediction \hat{Y}_t of the next outcome $Y_t \in [0, 1]$ of a bounded stationary ergodic process $(Y_t)_{t=-\infty, \dots, \infty}$ with knowledge of the past observations Y_1, \dots, Y_{t-1} . To evaluate the performance, a convex and M -lipschitz loss function $\ell : [0, 1]^2 \rightarrow [0, 1]$ is considered. The following fundamental limit has been proven by [Alg94]. For any prediction strategy, almost surely

$$\liminf_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \geq L^*, \text{ where } L^* = \mathbb{E} \left[\inf_{f \in \mathcal{B}^\infty} \mathbb{E} \left[\ell(f(Y_{-\infty}^{-1}), Y_0) | Y_{-\infty}^{-1} \right] \right] \quad (1)$$

is the expected minimal loss over all possible Borel estimations of the outcome Y_0 based on the infinite past (\mathcal{B}^∞ denotes the set of Borel functions from $[0, 1]^\infty$ to $[0, 1]$). One may thus try to design *consistent* strategies that achieve the lower bound, that is, $\limsup_T \{ (1/T) \sum_t \ell(\hat{Y}_t, Y_t) \} \leq L^*$.

Litterature review. Many forecasting strategies have been designed to this purpose. The vast majority of these strategies are based on statistical techniques used for time-series prediction, going from parametric models like autoregressive models (see [BD91]) to non-parametric methods (see the reviews of [GHSV89, Bos96, MF98]). In recent years, another collection of algorithms resolving related problems have been designed in [GLF01, GO07, BBGO10, BP11]. At their cores, all these algorithms use some machine learning non-parametric prediction scheme (like histogram, kernel, or nearest neighbor estimation) with

parameters by given both a window, and the length of the past to consider. Then, they output predictions by mixing the countably infinite set of experts corresponding to strategies with fixed values of these two parameters.

Our approach. We adopt the point of view of individual sequences, see the monograph of [CBL06]. In the process, we divide into two separate layers the setting of stochastic time series and the one of individual sequences. Our main result is Theorem 3 and it states that any strategy that satisfies some deterministic regret bound is consistent. Section 2 and 3 design such a strategy and consider the following framework of sequential prediction of individual sequences. We suppose that a sequence $(\mathbf{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ is observed step by step, where $\mathcal{X} \subset [0, 1]^d$ is the covariable space and $\mathcal{Y} \subset [0, 1]$ a convex observation space (in Section 3, \mathbf{x}_t will be replaced by $y_{t-d}^{t-1} = y_{t-d}, \dots, y_{t-1}$, then, y_t will be replaced by Y_t in Section 4). The learner is asked at each time step t to predict the next observation y_t with knowledge of the past observations y_1, \dots, y_{t-1} and of the past and present exogenous variables $\mathbf{x}_1, \dots, \mathbf{x}_t$. The goal of the forecaster is to minimize its cumulative regret against the class \mathcal{L}_L^d of L -Lipschitz functions from $[0, 1]^d$ to $[0, 1]$,

$$\widehat{R}_{L,T} = \sum_{t=1}^T \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^T \ell(f(\mathbf{x}_t), y_t),$$

that is, to ensure $\widehat{R}_{L,T} = o(T)$. In Section 2, we describe the nested EG strategy (Algorithm 2), which follows the spirit of binary regression trees like Cart (see [BFSO84]). We provide in Theorem 1 a finite-time regret bound with respect to the class of L -Lipschitz functions. We recall below the considered setting.

At each time step $t = 1, \dots, T$,

1. Forecaster observes $\mathbf{x}_t \in \mathcal{X} \subset [0, 1]^d$
2. Forecaster predicts $\widehat{y}_t \in [0, 1]$
3. Environment chooses $y_t \in \mathcal{Y}$
4. Forecaster suffers loss $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t) \in [0, 1]$.

Contributions. First, we clean up the standard analysis of prediction of ergodic processes by carrying out the aforementioned separation in two layers. The second advantage is the computational efficiency as we will discuss later in remarks. A third benefit of our approach is to be valid for a general class of loss functions when previous papers to our knowledge only treat particular cases like the square loss or the pinball loss.

2 The nested EG strategy

The nested EG strategy (Algorithm 2) incrementally builds an estimate of the best Lipschitz function f^* . The core idea is to estimate f^* precisely in areas of the covariable space \mathcal{X} with many occurrences of covariables \mathbf{x}_t , while estimating it loosely in other parts of the space. To implement this idea, Algorithm 2 maintains a deterministic binary tree whose nodes are associated with regions of

Parameter: $M > 0$

For time step $t = 1, 2, \dots$

1. Define the learning parameter $\eta_t = M^{-1} \sqrt{(\log 2)/t}$
2. Predict

$$\hat{y}_t = \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \ell'(\hat{y}_s, y_s))}{1 + \exp(-\eta_t \sum_{s=1}^{t-1} \ell'(\hat{y}_s, y_s))} \in [0, 1],$$

where ℓ' denotes the (sub)gradient of ℓ with respect to its first argument

3. Observe y_t
-

Algorithm 1: The gradient-based exponentially weighted average forecaster (EG) with two constant experts that predict respectively 0 and 1.

the covariable space, such that the regions with nodes deeper in the tree (further away from the root) represent increasingly smaller subsets of \mathcal{X} (see Figure 1).

In the later, we assume for simplicity that $\mathcal{X} = [0, 1]^d$ and $\mathcal{Y} = [0, 1]$ and that the loss function ℓ is from $[0, 1]^2$ to $[0, 1]$. The case of unknown bounded sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ will be treated later in remarks.

2.1 The best constant oracle

If the number of observations such that \mathbf{x}_t belong to a subset $\mathcal{X}^{\text{node}} \subset \mathcal{X}$ is small enough, one does not need to estimate f^* precisely over $\mathcal{X}^{\text{node}}$. Lemma 1 formalizes this idea by controlling the approximation error suffered by approximating f^* by the best constant in $[0, 1]$. The control is expressed in terms of the number of observations T^{node} and of the size of the set $\mathcal{X}^{\text{node}}$, which is measured by its diameter defined as $\text{diam}(\mathcal{X}^{\text{node}}) = \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^{\text{node}}} \|\mathbf{x} - \mathbf{x}'\|_2$.

Lemma 1 (Approximation of f^* by a constant). *Let $T^{\text{node}} \geq 1$ and suppose that ℓ is M -Lipschitz in its first argument. Then,*

$$\inf_{y \in [0, 1]} \sum_{t=1}^{T^{\text{node}}} \ell(y, y_t) \leq \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T^{\text{node}}} \ell(f(\mathbf{x}_t), y_t) + ML T^{\text{node}} \text{diam}(\mathcal{X}^{\text{node}}),$$

where $\mathcal{X}^{\text{node}} \subset [0, 1]^d$ is such that $\mathbf{x}_t \in \mathcal{X}^{\text{node}}$ for all $t = 1, \dots, T^{\text{node}}$.

Proof. Let $t \geq 1$. Using that ℓ is M -Lipschitz and f is L -Lipschitz, we get

$$\ell(f(\mathbf{x}_1), y_t) - \ell(f(\mathbf{x}_t), y_t) \leq M |f(\mathbf{x}_1) - f(\mathbf{x}_t)| \leq ML \|\mathbf{x}_1 - \mathbf{x}_t\|_2 \leq ML \delta.$$

Summing over t and noting that $\inf_y \sum_t \ell(y, y_t) \leq \sum_t \ell(f(\mathbf{x}_1), y_t)$ concludes. \square

2.2 Performing as well as the best constant: the EG strategy

Lemma 1 implies that considering constant predictions is not bad when either the covariable region is small, or the number of observations is small. The next step consists thus of estimating online the best constant prediction in $[0, 1]$.

To do so, among many existing methods, we consider the well-known *gradient-based exponentially weighted average forecaster* (EG), introduced by [KW97]. In the setting of prediction of individual sequences with expert advice—see the monograph by [CBL06], EG competes with the best fixed convex combination of experts. In the case where two experts predict constant predictions respectively 0 and 1 at all time steps, EG ensures vanishing average regret with respect to any constant prediction in $[0, 1]$. We describe in Algorithm 1 this particular case of EG and we provide the associated regret bound in Lemma 2, whose proof follows from the standard proof of EG, available for instance in [CBL06].

Lemma 2 (EG). *Let $T^{\text{node}} \geq 1$. We assume that the loss function ℓ is convex and M -Lipschitz in its first argument. Then, the cumulative loss of Algorithm 1 is upper bounded as follows:*

$$\sum_{t=1}^{T^{\text{node}}} \ell(\hat{y}_t, y_t) \leq \inf_{y \in [0,1]} \sum_{t=1}^{T^{\text{node}}} \ell(y, y_t) + 2M\sqrt{T^{\text{node}} \log 2}.$$

Unknown value of M . Note that Algorithm 1 needs to know in advance a uniform bound M on ℓ' . This is the case, if one considers as we do a bounded observation space $[0, 1]$ with the absolute loss function, defined for all $y, y' \in [0, 1]$ by $\ell(y', y) = |y - y'|$; the pinball loss, defined by $\ell_\alpha(y', y) = (\alpha - \mathbf{1}_{\{y \geq x\}})(y - y')$; or the square loss, defined by $\ell(y', y) = (y - y')^2$. However, in the case of an unknown observation space \mathcal{Y} the bound on the gradient of the square loss is unknown and needs to be calibrated online at the small cost of the additional term $2M(2 + 4(\log 2)/3)$ in the regret bound, see [dRvEGK14].

2.3 The nested EG strategy

The nested EG strategy presented in Algorithm 2 implements the idea of Lemma 1 and Lemma 2. It maintains a binary tree whose nodes are associated with regions of the covariable space $[0, 1]^d$. The nodes in the tree are indexed by pairs of integers (h, i) ; where the first index $h \geq 0$ denotes the distance of the node to the root (also referred to as the depth of the node) and the second index i belongs

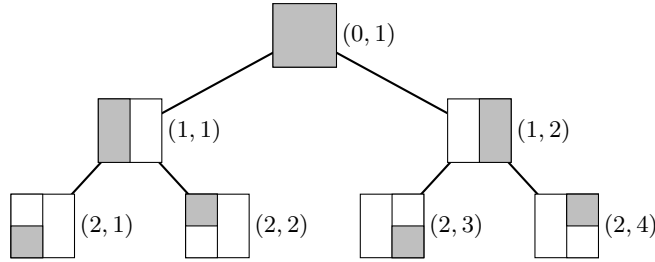


Fig. 1. Representation of the binary tree in dimension $d = 2$.

to $\{1, \dots, 2^h\}$. The root is thus denoted by $(0, 1)$. By convention, $(h+1, 2i-1)$ and $(h+1, 2i)$ are used to refer to the two children of node (h, i) . Let $\mathcal{X}^{(h,i)}$ be the region associated with node (h, i) . By assumption, these regions are hyper-rectangle and must satisfy the constraints

$$\mathcal{X}^{(0,1)} = [0, 1]^d \quad \text{and} \quad \mathcal{X}^{(h,i)} = \mathcal{X}^{(h+1,2i-1)} \sqcup \mathcal{X}^{(h+1,2i)},$$

where \sqcup denotes the disjoint union. The set of regions associated with terminal nodes (or leaves) forms thus a partition of $[0, 1]^d$.

At time step t , when a new covariable \mathbf{x}_t is observed, Algorithm 2 first selects the associated leaf (h_t, i_t) such that $\mathbf{x}_t \in \mathcal{X}^{(h_t, i_t)}$ (step 2). The leaf (h_t, i_t) then predicts the next observation y_t by updating a local version $\mathcal{E}^{(h_t, i_t)}$ of Algorithm 1 (step 3). Namely, $\mathcal{E}^{(h_t, i_t)}$ runs Algorithm 1 on the sub-sequence of observations (\mathbf{x}_s, y_s) such that the associated leaf is (h_t, i_t) , that is, $(h_s, i_s) = (h_t, i_t)$. When the number of observations $T^{(h_t, i_t)}$ received and predicted by leaf (h_t, i_t) becomes too large compared to the size of the region $\mathcal{X}^{(h_t, i_t)}$ (step 6), the tree is updated. To do so, the region $\mathcal{X}^{(h_t, i_t)}$ is divided in two sub-regions of equal volume by cutting along one given coordinate.

The coordinate $r_t + 1$ to be split is chosen in a deterministic order, where $r_t = (h_t \bmod d)$ and \bmod denotes the modulo operation. Thus, at the root node $(0, 1)$ the first coordinate is split, then by going down in the tree we split the second one, then the third one and so on until we reach the depth d , in which case we split the first coordinate for the second time. Each sub-region is associated with a child of node (h_t, i_t) . Consequently, (h_t, i_t) becomes an inner node and is thus no longer used to form predictions.

To facilitate the formal study of the algorithm, we will need some additional notation. In particular, we will introduce time-indexed versions of several quantities. \mathcal{T}_t denotes the tree stored by Algorithm 2 at the beginning of time step t . The initial tree is thus the root $\mathcal{T}_0 = \{(0, 1)\}$ and it is expanded when the splitting condition (step 6) holds, as

$$\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{(h_t + 1, 2i_t - 1), (h_t + 1, 2i_t)\}$$

(step 6.3) and remains unchanged otherwise. We denote by N_t the number of nodes of \mathcal{T}_t and by H_t the height of \mathcal{T}_t , that is, the maximal depth of the leaves of \mathcal{T}_t . A performance bound for Algorithm 2 is provided below.

Theorem 1. *Let $T \geq 1$ and $d \geq 1$. Then, the cumulative regret $\hat{R}_{L,T}$ of Algorithm 2 is upper bounded as*

$$\begin{aligned} \sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^T \ell(f(\mathbf{x}_t), y_t) &\leq M(3+L) \sqrt{N_T T} \\ &\leq M(3+L) \left(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}} \right). \end{aligned}$$

Time and storage complexity. The following lemma provides time and storage complexity guarantees for Algorithm 2. It upper bounds the maximal size of \mathcal{T}_T , that is, its number of nodes N_T and its depth H_T , which yields in particular the regret bound of order $O(T^{(d+1)/(d+2)})$ stated in Theorem 1.

Initialization:

- $\mathcal{T} = \{(0, 1)\}$ a tree (for now reduced at a root node)
- Define the bin $\mathcal{X}^{(0,1)} = [0, 1]^d$
- Start $\mathcal{E}^{(0,1)}$ a replicate of Algorithm 1

For $t = 1, \dots, T$

1. Observe $\mathbf{x}_t \in [0, 1]^d$
2. Select the leaf (h_t, i_t) such that $\mathbf{x}_t \in \mathcal{X}^{(h_t, i_t)}$
3. Predict according to $\mathcal{E}^{(h_t, i_t)}$
4. Observe y_t and feed $\mathcal{E}^{(h_t, i_t)}$ with it
5. Update the number of observations predicted by $\mathcal{E}^{(h_t, i_t)}$

$$T^{(h_t, i_t)} \leftarrow \#\{1 \leq s \leq t, (h_s, i_s) = (h_t, i_t)\}$$
6. **If** the splitting condition $T^{(h_t, i_t)} + 1 \geq \left(\text{diam}(\mathcal{X}^{(h_t, i_t)})\right)^{-2}$ **holds then** extend the binary tree \mathcal{T} as follows:
 - 6.1. Compute the decomposition $h_t = k_t d + r_t$ with $r_t \in \{0, \dots, d-1\}$
 - 6.2. Split coordinate $r_t + 1$ for node (h_t, i_t)
 - 6.2.1. Define the splitting threshold $\tau = (x^- + x^+)/2$, where

$$x^- = \inf_{\mathbf{x} \in \mathcal{X}^{(h_t, i_t)}} \{x_{r_t+1}\} \text{ and } x^+ = \sup_{\mathbf{x} \in \mathcal{X}^{(h_t, i_t)}} \{x_{r_t+1}\}.$$
 - 6.2.2. Define two children leaves for node (h_t, i_t) :
 - the left leaf $(h_t + 1, 2i_t - 1)$ with corresponding bin

$$\mathcal{X}^{(h_t+1, 2i_t-1)} = \{\mathbf{x} \in \mathcal{X}^{(h_t, i_t)} : x_{r_t+1} \in [x^-, \tau]\}$$
 - the right leaf $(h_t + 1, 2i_t)$ with corresponding bin

$$\mathcal{X}^{(h_t+1, 2i_t)} = \left\{ \mathbf{x} \in \mathcal{X}^{(h_t, i_t)} : \begin{array}{l} x_{r_t+1} \in [\tau, x^+ \text{ if } x_+ < 1 \\ x_{r_t+1} \in [\tau, 1] \text{ if } x_+ = 1 \end{array} \right\}$$
 - 6.2.3. Update $\mathcal{T} \leftarrow \mathcal{T} \cup \{(h_t + 1, 2i_t - 1), (h_t + 1, 2i_t)\}$

Algorithm 2: Sequential prediction of function via Nested EG

Lemma 3. *Let $T \geq 1$ and $d \geq 1$. Then the depth H_T and the number of nodes N_T of the binary tree \mathcal{T}_T stored by Algorithm 2 after T time steps are upper bounded as follows:*

$$H_T \leq 1 + \frac{d}{2} \log_2(4dT) \quad \text{and} \quad N_T \leq 1 + 8(dT)^{\frac{d}{d+2}}.$$

Indeed, Algorithm 2 needs to store a constant number of parameters at each node of the tree. Thus the space complexity is of order $O(N_T) = O(T^{d/(d+2)})$. Besides at each time step t , Algorithm 2 needs to perform $O(H_t) = O(\log t)$ binary test operations in order to select the leaf (h_t, i_t) . It then only needs constant time to update both $\mathcal{E}^{(h_t, i_t)}$ and \mathcal{T} . Thus the per-round time complexity of Algorithm 2 is of order $O(\log t)$ and the global time complexity is of order $O(T \log T)$. Therefore, we can summarize:

$$\text{Storage complexity: } O(T^{d/(d+2)}), \quad \text{Time complexity: } O(T \log T).$$

Unknown bounded sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. As we mentioned in the end of Section 2.2, the generalization of Algorithm 1 and thus of Algorithm 2 to

an unknown set $\mathcal{Y} \subset \mathbb{R}$ can be obtained by using standard tools of individual sequences—see for instance [dRvEGK14]. To adapt Algorithm 2 to any unknown compact set $\mathcal{X} \subset \mathbb{R}^d$, one can first divide the covariable space \mathbb{R}^d in hyper-rectangle subregions of the form $[n_1, n_1 + 1] \times \cdots \times [n_d, n_d + 1]$ and then run independent versions of Algorithm 2 on all of these subregions. If $\text{diam}(\mathcal{X}) \leq \sqrt{d}B$ with an unknown value of $B > 0$, then the number of initial subregions is upper-bounded by $\lceil B \rceil^d$ and by Jensen’s inequality, this adaptation would lead to a multiplicative cost of $\lceil B \rceil^{d/(d+2)}$ in the upper-bound of Theorem 1.

Comparison with other methods. One may want to obtain similar guarantees by considering other strategies like uniform histograms, kernel regression, or nearest neighbors, which were studied in the context of stationary ergodic processes by [GLF01,GO07,BBGO10,BP11]. We were unfortunately unable to provide any finite-time and deterministic analysis neither for kernel regression nor for nearest neighbors estimation. The regret bound of Theorem 1 can however be obtained in an easier manner with uniform histograms. To do so, one can consider the class of uniform histograms \mathcal{H}_N . We divide the covariable space $[0, 1]^d$ in a partition $(I_j)_{j=1,\dots,N}$ of N subregions of equal size. We define \mathcal{H}_N as the class of 2^N prediction strategies that predict the constant values 0 or 1 in each bin of the partition. Competing with this class \mathcal{H}_N of 2^N functions by resorting for instance to EG gives the regret bound

$$\sum_{t=1}^T \ell(\hat{y}_t, y_t) \leq \min_{z \in [0,1]^N} \sum_{t=1}^T \ell\left(\sum_{j=1}^N z_j \mathbf{1}_{I_j}(\mathbf{x}_t), y_t\right) + 2M\sqrt{TN}.$$

Now, optimizing the number N of bins in hindsight (or by resorting to the doubling trick) provides a regret bound of order $O(T^{(d+1)/(d+2)})$ against any Lipschitz function. The size of the class \mathcal{H}_N is however exponential in $N = O(T^{d/(d+2)})$, which makes the method computationally inefficient.

However, in the worst case the nested EG strategy has no better guarantee. Such worst case occurs for large number N_T of nodes, which happens in particular when the trees are height-balanced, that is, when the covariables \mathbf{x}_t are uniformly distributed in $[0, 1]^d$. But the nested EG strategy adapts better to data. If the covariables \mathbf{x}_t are non-uniformly allocated (with regions of the space $[0, 1]^d$ associated with much more observations than in other regions of similar size), the resulting tree \mathcal{T}_T will be un-balanced, leading to a smaller number of nodes. In the best case, $N_T = O(H_T)$, which yields a regret of order $O(\sqrt{T \log T})$. By improving the definition of Algorithm 2, one can even obtain the optimal and expected $O(\sqrt{T})$ regret if (\mathbf{x}_t) is constant. To do so, it only needs to compute online the effective range of the data that belongs to each node (h, i) ,

$$\delta_t^{(h,i)} = \text{diam} \{ \mathbf{x}_s, \quad 0 \leq s \leq t \text{ and } (h_s, i_s) = (h, i) \}$$

and substitute the diameter $\text{diam } \mathcal{X}^{(h,i)}$ by $\delta_{t+1}^{(h,i)}$ in the splitting condition of the algorithm (step 6).

Proofs. The proofs of Theorem 1 and Lemma 3 are based on the following lemma, which controls the size of the regions associated with nodes located at depth h in the tree \mathcal{T}_T .

Lemma 4. *Let $h \geq 0$. Then, for all indices $i = 1, \dots, 2^h$, the diameter of the region $\mathcal{X}^{(h,i)}$ associated with node (h,i) in Algorithm 2 is upper bounded as*

$$p \operatorname{diam}(\mathcal{X}^{(h,i)}) \leq \sqrt{2d} 2^{-h/d}.$$

Basically, the proof of Lemma 4 consists of an induction on the depth h . It is postponed to Appendix A.

Proof (of Lemma 3). **Upper bound for N_T .** For each node (h,i) , we recall that $T^{(h,i)} = \sum_{t=1}^T \mathbb{1}_{\{(h,i_t)=(h,i)\}}$ denotes the number of observations predicted by using algorithm $\mathcal{E}^{(h,i)}$. The total number of observations T is the sum of $T^{(h,i)}$ over all nodes (h,i) . That is,

$$T = \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} T^{(h,i)} \mathbb{1}_{\{(h,i) \in \mathcal{T}_T\}} \geq \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} T^{(h,i)} \mathbb{1}_{\{(h,i) \text{ is an inner node in } \mathcal{T}_T\}}.$$

Now we use the fact that each inner node (h,i) has reached its splitting condition (step 6 of Algorithm 2), that is, $T^{(h,i)} + 1 \geq (\operatorname{diam}(\mathcal{X}^{(h,i)}))^{-2}$. Using that $\operatorname{diam}(\mathcal{X}^{(h,i)}) \leq \sqrt{2d} 2^{-h/d}$ by Lemma 4, we get

$$\begin{aligned} T &\geq \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} \left[-1 + \left(\operatorname{diam}(\mathcal{X}^{(h,i)}) \right)^{-2} \right] \mathbb{1}_{\{(h,i) \text{ is an inner node}\}} \\ &\geq \sum_{h=0}^{H_T} \underbrace{\left(-1 + \frac{2^{2h/d}}{2d} \right)}_{g(h)} \underbrace{\sum_{i=1}^{2^h} \mathbb{1}_{\{(h,i) \text{ is an inner node}\}}}_{n_h}. \end{aligned} \quad (2)$$

Because $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ is convex in h , by Jensen's inequality

$$T \geq N_T^{\text{in}} g\left(\frac{1}{N_T^{\text{in}}} \sum_{h=0}^{H_T} h n_h\right),$$

where $N_T^{\text{in}} = \sum_h n_h$ is the total number of inner nodes. Now, by Lemma 8 in Appendix B, because \mathcal{T}_T is a binary tree with N_T nodes in total, it has exactly $N_T^{\text{in}} = (N_T - 1)/2$ inner nodes and the average depth of its inner nodes is lower-bounded as

$$\frac{1}{N_T^{\text{in}}} \sum_{h=0}^{H_T} h n_h \geq \log_2\left(\frac{N_T - 1}{8}\right).$$

Substituting in the previous bound, it implies

$$\begin{aligned} T &\geq \frac{N_T - 1}{2} g\left(\log_2\left(\frac{N_T - 1}{8}\right)\right) = \frac{N_T - 1}{2} \left(-1 + \frac{1}{2d} 2^{\frac{2}{d} \log_2((N_T - 1)/8)} \right) \\ &= -\frac{N_T - 1}{2} + \frac{N_T - 1}{4d} \left(\frac{N_T - 1}{8} \right)^{2/d} \geq \underbrace{-\frac{N_T - 1}{2}}_{\geq -T/2} + \frac{2}{d} \left(\frac{N_T - 1}{8} \right)^{1+2/d}. \end{aligned}$$

By reorganizing the terms, it entails $dT \geq (3/4)dT \geq ((N_T - 1)/8)^{1+2/d}$. Thus, $(N_T - 1)/8 \leq (dT)^{d/(d+2)}$, which yields the desired bound for N_T .

Upper bound for H_T . We start from (2) and we use the fact that for all $h = 0, \dots, H_T - 1$, there exists at least one inner node of depth h in \mathcal{T} . Thus,

$$T \geq \sum_{h=0}^{H_T-1} \left(-1 + \frac{2^{2h/d}}{2d} \right) = -H_T + \frac{1}{2d} \frac{2^{2H_T/d} - 1}{2^{2/d} - 1} \geq -H_T + \frac{2^{2(H_T-1)/d}}{2d}$$

where the last inequality is because $(a-1)/(b-1) \geq a/b$ for all numbers $a \geq b > 1$. Therefore, by upper-bounding $T \geq H_T$, we get $4T \geq 2^{2(H_T-1)/d}/d$ and thus $2(H_T-1)/d \leq \log_2(4dT)$ which concludes the proof. \square

Proof (of Theorem 1). The cumulative regret suffered by Algorithm 2 is controlled by the sum of all cumulative regrets incurred by algorithms $\mathcal{E}^{(h,i)}$. That is,

$$\widehat{R}_{L,T} \leq \sum_{(h,i) \in \mathcal{T}_T} \left[\sum_{t \in S^{(h,i)}} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t \in S^{(h,i)}} \ell(f(\mathbf{x}_t), y_t) \right],$$

where $S^{(h,i)} = \{1 \leq t \leq T : (h_t, i_t) = (h, i)\}$ is the set of time steps assigned to node (h, i) . Now, by Lemma 2, the cumulative loss incurred by $\mathcal{E}^{(h,i)}$ satisfies

$$\begin{aligned} \sum_{t \in S^{(h,i)}} \ell(\widehat{y}_t, y_t) &\leq \inf_{y \in [0,1]} \sum_{t \in S^{(h,i)}} \ell(y, y_t) + 2M \sqrt{T^{(h,i)} \log 2} \\ &\leq \inf_{f \in \mathcal{L}_L^d} \sum_{t \in S^{(h,i)}} \ell(f(\mathbf{x}_t), y_t) + \underbrace{ML \operatorname{diam}(\mathcal{X}^{(h,i)})}_{\leq 1/\sqrt{T^{(h,i)}}} T^{(h,i)} + 2M \sqrt{T^{(h,i)} \log 2} \\ &\leq 1/\sqrt{T^{(h,i)}} \text{ by step 6 of Algorithm 2} \end{aligned}$$

where the second inequality is by Lemma 1. Thus,

$$\widehat{R}_{L,T} \leq M \left(L + \underbrace{2\sqrt{\log 2}}_{\leq 3} \right) \sum_{(h,i) \in \mathcal{T}_T} \sqrt{T^{(h,i)}}.$$

Then, by Jensen's inequality,

$$\frac{1}{N_T} \sum_{(h,i) \in \mathcal{T}_T} \sqrt{T^{(h,i)}} \leq \sqrt{\frac{1}{N_T} \sum_{(h,i)} T^{(h,i)}} = \sqrt{\frac{T}{N_T}},$$

which concludes the first statement of the theorem. The second statement follows from Lemma 3 and because for all $a, b \geq 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\begin{aligned} M(3+L)\sqrt{N_T T} &\leq M(3+L)\sqrt{\left(1 + 4(3dT)^{d/(d+2)}\right)T} \\ &\leq M(3+L)\left(\sqrt{T} + \sqrt{4(3dT)^{d/(d+2)}T}\right) = M(3+L)\left(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}}T^{\frac{d+1}{d+2}}\right). \end{aligned}$$

\square

3 Autoregressive framework

We present in this section a technical result that will be useful for later purposes. Here, the forecaster still sequentially observes from time $t = 1$ an arbitrary bounded sequence $(y_t)_{t=-\infty, \dots, +\infty}$. However, at time step t , it is asked to forecast the next outcome $y_t \in [0, 1]$ with knowledge of the past observations $y_1^{t-1} = y_1, \dots, y_{t-1}$ only. We are interested in a strategy that performs asymptotically as well as the best model that considers the last d observations to form the predictions, and this simultaneously for all values of $d \geq 1$. More formally, we denote

$$\hat{R}_{L,T}^d \triangleq \sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^T \ell(f(y_{t-d}^{t-1}), y_t),$$

and we want that for all d , the average regrets $\hat{R}_{L,T}^d/T$ vanish as $T \rightarrow \infty$. We show how it can be obtained via a meta-algorithm (Algorithm 4) that combines an increasing sequence of nested EG forecasters described in Algorithm 3. The sequence is denoted by $\mathcal{A}_1, \mathcal{A}_2, \dots$ and is such that for each $d \geq 1$, \mathcal{A}_d^\dagger forms predictions for $t \geq t_d$ for some starting time $t_d \geq 1$ and satisfies the regret bound stated in Lemma 5.

Parameter: $d \geq 1$ and t_d , a starting time

For $t \leq t_d - 1$

Form no prediction[†] and observe y_t

For $t = t_d, \dots, T$

1. define $\mathbf{x}_t = y_{t-d}^{t-1}$ and feed Algorithm 2 with $\mathbf{x}_t \in [0, 1]^d$
 2. predict $f_{d,t}$ according to Algorithm 2 and feed Algorithm 2 with y_t
-

Algorithm 3: Forecaster \mathcal{A}_d for fixed past d .

Lemma 5 (Fixed past d). *Let $T \geq 1$, $d \geq 1$, $L > 0$, and $t_d \geq d + 1$. Then, Algorithm 3 has a regret upper-bounded as*

$$\sum_{t=t_d}^T \ell(f_{d,t}, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=t_d}^T \ell(f(y_{t-d}^{t-1}), y_t) \leq M(3 + L) \left(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}} \right).$$

Proof. The regret bound is a straightforward corollary of Theorem 1. \square

Now we show how to obtain the regret bound of Lemma 5 simultaneously for all $d \geq 1$. To do so, we consider an increasing sequence of integers (t_d) such that $t_1 = 2$. Namely, t_d states at which time step algorithm \mathcal{A}_d starts to form

[†] Algorithm \mathcal{A}_d will only be used by a meta-algorithm for time steps $t \geq t_d$

Parameter:

- (t_d) an increasing sequence of starting times
- $(\mathcal{F}_d)_{d \geq 1}$ a sequence of forecasters such that \mathcal{F}_d forms predictions for time steps $t \geq t_d$
- (η_t) a sequence of learning rates

Initialization:

- **For** $t = 1, \dots, t_1 - 1$, predict $\hat{y}_t = 1/2$
- set $D_{t_1} = 1$ and $\hat{p}_{1,t_1} = 1$

For $t = t_1, \dots, T$

1. **For** each $d = 1, \dots, D_t$, denote by $f_{d,t}$ the prediction formed by \mathcal{F}_d
2. predict $\hat{y}_t = \sum_{d=1}^{D_t} \hat{p}_{d,t} f_{d,t}$
3. update the number of active forecasters
 - 3.1 if the next starting time occurs in $t + 1$, i.e., $t_{D_{t+1}} = t + 1$ then
 - increase the number of forecasters by 1: $D_{t+1} = D_t + 1$
 - initialize the weight of the new forecaster: $p_{D_{t+1},t+1} = 1/D_{t+1}$
 - 3.2 otherwise if no expert starts in $t + 1$, make no change: $D_{t+1} = D_t$
4. observe Y_t and perform exponential weight update component-wise for $d = 1, \dots, D_t$ as

$$\hat{p}_{d,t+1} = \frac{D_t}{D_{t+1}} \frac{\hat{p}_{d,t}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell(f_{d,t}, y_t)}}{\sum_{k=1}^{D_t} \hat{p}_{k,t}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell(f_{k,t}, y_t)}}.$$

Algorithm 4: Extension of the Algorithm 2 to unknown past d .

predictions and thus to be combined in Algorithm 4. We define at each time step $s \geq 1$ the number of active algorithms $D_s = \sup\{d \geq 1 : t_d \leq s\}$. Basically, Algorithm 4 is a meta-algorithm that combines via EG the predictions formed by all forecasters \mathcal{A}_d for $d \geq 1$. Note that at time step t , only the D_t first forecasters $\mathcal{A}_1, \dots, \mathcal{A}_{D_t}$ suggest predictions.

Lemma 6 controls the cumulative loss of Algorithm 4 by the cumulative loss of the best strategy \mathcal{F}_d . The comparison is performed only on the time steps where \mathcal{F}_d is active (i.e., forms a prediction).

Lemma 6. *Let $T \geq 1$ and $(\eta_t)_{t \geq 1}$ be a decreasing sequence of non-negative learning rates. Then, Algorithm 4 satisfies for all $d \in 1, \dots, D_T \triangleq \sup\{d, t_d \leq T\}$*

$$\sum_{t=t_d}^T \ell(\hat{y}_t, y_t) - \ell(f_{d,t}, y_t) \leq \frac{1}{\eta_{T+1}} \log(D_{T+1}) + \frac{1}{8} \sum_{t=t_d}^T \eta_t,$$

which implies with learning rates $\eta_t = 2/\sqrt{t}$ for $t \geq 1$ the following regret bound

$$\sum_{t=t_d}^T \ell(\hat{y}_t, y_t) - \ell(f_{d,t}, y_t) \leq \sqrt{T+1} \log D_{T+1}.$$

Note that the choice $\eta_t = \min_{s \leq t} \sqrt{\log D_t/t}$ for $t \geq 1$ may yield the right dependency $\sqrt{\log D_T}$ in the number of experts. Similarly, the term \sqrt{T} can be replaced by $\sqrt{T - t_d + 1}$ by considering for instance the aggregation rule of [GSvE14] with one learning rate sequence for each expert. The proof of Lemma 6 follows the standard one of the exponentially weighted average forecaster. It is postponed to Appendix C. It could also be recovered by noting that our setting with starting experts is almost a particular case of the setting of sleeping experts introduced in [FSSW97]. We could thus obtain similar results by following algorithms and proofs designed for this setting. We write “almost” because here we do not know in advance the final number of active experts, which explains the non-optimal term in D_t .

Theorem 2. *Let $T \geq 1$, $L > 0$. Let (t_d) be an increasing sequence of integers such that $t_1 = 2$. Then, for all $d \leq D_T \triangleq \sup\{d, t_d \leq T\}$, Algorithm 4 run with an increasing sequence (t_d) of starting times, sequence of forecasters (\mathcal{A}_d) and sequence of learning rates $\eta_t = 2/\sqrt{t}$ satisfies*

$$\begin{aligned} \widehat{R}_{L,T}^d &= \sum_{t=1}^T \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^T \ell(f(y_{t-d}^{t-1}), y_t) \\ &\leq t_d + \sqrt{T+1} \log D_{T+1} + M(3+L) \left(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}} \right). \end{aligned}$$

Consequently, for all $d \geq 1$, $\limsup_{T \rightarrow \infty} \left(\widehat{R}_{L,T}^d / T \right) \leq 0$.

Proof. The regret bound is by combining Lemma 5 and Lemma 6, together with $\ell(\widehat{y}_t, y_t) \leq 1$ for $t < t_d$. The second part is obtained by dividing by T and making T grows to infinity. The last part is then a consequence of Theorem 2. \square

4 Convergence to L^*

In this section, we present our main result by deriving from Theorem 2 similar results obtained in a stochastic setting by [GLF01,GO07,BBGO10,BP11].

We leave here the setting of individual sequences of the previous sections and we assume that the sequence of observations y_1, \dots, y_T is now generated by some stationary ergodic process. More formally, we assume that a stationary bounded ergodic process $(Y_t)_{t=-\infty, \dots, \infty}$ is sequentially observed. At time step t , the learner is asked to form a prediction \widehat{Y}_t of the next outcome $Y_t \in [0, 1]$ of the sequence with knowledge of the past observations $Y_1^{t-1} = Y_1, \dots, Y_{t-1}$. The nested EG strategy, as a consequence of the deterministic regret bound of Theorem 1, will be shown to be consistent. We recall that [Alg94] proved that all prediction strategies verify almost surely $\liminf_{T \rightarrow \infty} \left\{ \frac{1}{T} \sum_{t=1}^T \ell(\widehat{Y}_t, Y_t) \right\} \geq L^*$, where L^* , defined in (1), is the expected minimal loss over all possible Borel estimations of the outcome Y_0 based on the infinite past. To put it another way: we cannot hope to design strategies outperforming L^* . It is thus usual to require that $\sum_{t=1}^T \ell(\widehat{Y}_t, Y_t)/T$ tends to L^* as $T \rightarrow \infty$.

From individual sequences to ergodic processes

Theorem 3 shows that any strategy that achieves a deterministic regret bound for individual sequences as in Theorem 2 predicts asymptotically as well as the best strategy defined by a Borel function.

Theorem 3 will make two main assumptions on the ergodic sequence to be predicted. First, the sequence is supposed to lie in $[0, 1]$. As earlier, this assumption can be easily relaxed to any bounded subset of \mathbb{R} —see remarks of Sections 2.2 and 2.3. The generalization to unbounded sequence is left to future work and should follow from [GO07]. Second, Theorem 3 assumes that for all $d \geq 1$ the law of Y_{-d}^{-1} is regular, that is, for any Borel set $S \subset [0, 1]^d$ and for any $\varepsilon > 0$, one can find a compact set K and an open set V such that

$$K \subset S \subset V, \quad \text{and} \quad \mathbb{P}_{Y_{-d}^{-1}}(V \setminus K) \leq \varepsilon.$$

This second assumption is considerably weaker than the assumptions required by [BP11] on the law of (Y_{-d}^{-1}) obtained for quantile prediction. The authors indeed imposed that the random variables $\|Y_{-d}^{-1} - s\|$ have continuous distribution functions for all $s \in \mathbb{R}^d$ and the conditional distribution function $F_{Y_0|Y_{-\infty}^{-1}}$ to be increasing. One can however argue that their assumptions are thus hardly comparable with ours because they consider unbounded ergodic processes. We aim at obtaining in the future minimal assumptions for any generic convex loss function ℓ in the case of unbounded ergodic process, see [MW11].

Theorem 3. *Let $(Y_t)_{t=-\infty, \dots, \infty}$ be a stationary bounded ergodic process. We assume that for all t , $Y_t \in [0, 1]$ almost surely and that for all $d \geq 1$ the law of $Y_{-d}^{-1} = (Y_{-d}, \dots, Y_{-1})$ is regular. Let $\ell : [0, 1]^2 \rightarrow [0, 1]$ be a loss function M -Lipschitz in its first argument. Assume that a prediction strategy satisfies for all $d \geq 1$,*

$$\forall L \geq 0 \quad \limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \leq \limsup_{T \rightarrow \infty} \left(\inf_{f \in \mathcal{L}_L^d} \frac{1}{T} \sum_{t=1}^T \ell(f(Y_{t-d}^{-1}), Y_t) \right),$$

then, almost surely,

$$\limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \leq L^*.$$

By Theorem 2, Algorithm 4 satisfies the assumption of Theorem 3. Our deterministic strategy is thus asymptotically optimal for any stationary bounded ergodic process satisfying the assumptions of Theorem 3. Here we only give the main ideas in the proof of Theorem 3. The complete argument is given in Appendix E.

Proof (sketch for Theorem 3). The proof follows from the one of Theorem 1 in [GLF01]. The new ingredient of our proof is mainly Lemma 7, which states that the best constant Lipschitz strategy performs as well as the best constant Borel

strategy. First, because of Breiman's generalized ergodic theorem (see [Bre57]) the right-term converges, and by making $L \rightarrow \infty$, we get

$$\limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\widehat{Y}_t, Y_t) \right) \leq \inf_{f \in \mathcal{L}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)] ,$$

where \mathcal{L}^d is the set of Lipschitz functions from \mathbb{R}^d to \mathbb{R} . Then, by Lemma 7 the infimum over all Lipschitz functions equals the infimum over the set \mathcal{B}^d of Borel functions. Therefore, by exhibiting a well-chosen Borel function (see [Alg94, Theorem 8]), this yields

$$\begin{aligned} \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \ell(\widehat{Y}_t, Y_t) &\leq \inf_{f \in \mathcal{B}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)] \\ &= \mathbb{E} \left[\inf_{f \in \mathcal{B}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0) \mid Y_{-d}^{-1}] \right]. \end{aligned}$$

The proof is then concluded by making $d \rightarrow \infty$ thanks to the martingale convergence theorem. \square

Lemma 7. *Let \mathcal{X} be a convex and compact subset of a normed space. Let $\ell : [0, 1]^2 \rightarrow [0, 1]$ be a loss function M -Lipschitz in its first argument. Let X be a random variable on \mathcal{X} with a regular law \mathbb{P}_X and let Y be a random variable on $[0, 1]$. Then,*

$$\inf_{f \in \mathcal{L}^{\mathcal{X}}} \mathbb{E}[\ell(f(X), Y)] = \inf_{f \in \mathcal{B}^{\mathcal{X}}} \mathbb{E}[\ell(f(X), Y)] ,$$

where $\mathcal{L}^{\mathcal{X}}$ denotes the set of Lipschitz functions from \mathcal{X} to \mathbb{R} and $\mathcal{B}^{\mathcal{X}}$ the one of Borel functions from \mathcal{X} to \mathbb{R} .

The proof of Lemma 7 postponed to Appendix D as well. It follows from the Stone-Weierstrass theorem, used to approximate continuous functions, and from Lusin's theorem, to approximate Borel functions.

Computational efficiency. The space complexity of Algorithm 4 depends on the chosen sequence of starting times (t_d) . It can be arbitrary close to the space complexity of the nested EG strategy, which is $O(T^{d/(d+2)})$. Previous algorithms of [GLF01, GO07, BBGO10, BP11] exhibit consistent strategies as well. However, in practice, these algorithms involve choices of parameters somewhere in their design (by choosing the a priori weight of the infinite set of experts). Then, the consideration of an infinite set of experts makes the exact algorithm computationally inefficient. For practical purpose, it needs to be approximated. This can be obtained by MCMC or for instance by restricting the set of experts to some finite subset at the cost, however, of loosing theoretical guarantees, see [BP11].

Generic loss function. Theorem 3 assumes ℓ to be bounded, convex, and M -Lipschitz in its first argument. In contrast, the results of [GLF01, GO07, BBGO10] only hold for the square loss (while [BP11] extend them to the pinball-loss).

References

- Alg94. Paul H. Algoet. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, 40(3):609–633, 1994.
- BBGO10. Gérard Biau, Kevin Bleakley, László Györfi, and György Ottucsák. Non-parametric sequential prediction of time series. *Journal of Nonparametric Statistics*, 22(3):297–317, 2010.
- BD91. Peter J. Brockwell and Richard A. Davis. *Time series : theory and methods*. Springer Series in Statistics. Springer, New York, 1991.
- BFSO84. Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.
- Bos96. Denis Bosq. *Nonparametric statistics for stochastic processes : estimation and prediction*. Lecture notes in statistics. Springer, New York, 1996.
- BP11. Gérard Biau and Benoît Patra. Sequential quantile prediction of time series. *IEEE Transactions on Information Theory*, 57(3):1664–1674, 2011.
- Bre57. Leo Breiman. The individual ergodic theorem of information theory. *Annals of Mathematical Statistics*, 31:809–811, 1957.
- CBL06. Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Cho65. Yuan Shih Chow. Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics*, 36:552–558, 1965.
- dRvEGK14. Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.
- FSSW97. Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of STOC*, pages 334–343, 1997.
- Geo67. Georges Georganopoulos. Sur l’approximation des fonctions continues par des fonctions lipschitziennes. *Comptes Rendus de l’Académie des sciences*, 264(7):319–321, 1967.
- GHSV89. László Györfi, Wolfgang Härdle, Pascal Sarda, and Philippe Vieu. *Non-parametric curve estimation from time series*. Number 60 in Lecture notes in statistics. Springer-Verlag, Berlin, 1989.
- GLF01. László Györfi, Gábor Lugosi, and Ramon Trias Fargas. Strategies for sequential prediction of stationary time series, 2001.
- GO07. László Györfi and György Ottucsák. Sequential Prediction of Unbounded Stationary Time Series. *Information Theory, IEEE Transactions on*, 53(5):1866–1872, 2007.
- GSvE14. Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of COLT*, 2014.
- KW97. Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.
- MF98. Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.
- MW11. Gusztáv Morvai and Benjamin Weiss. Nonparametric sequential prediction for stationary processes. *Ann. Probab.*, 39(3):1137–1160, 2011.

- SL07. Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59:187–208, 2007.

Additional Material for “A consistent deterministic regression tree for non-parametric prediction of time series”

We gather in this appendix the proofs, which were omitted from the main body of the paper

A Proof of Lemma 4

It suffices to prove that for all $h \geq 0$, for all indexes $i \in \{1, \dots, 2^h\}$ and all coordinates $j \in \{1, \dots, d\}$, the ranges $\delta_j^{(h,i)} \triangleq \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^{(h,i)}} |x_j - x'_j|$ satisfies

$$\delta_j^{(h,i)} = \begin{cases} 2^{-(k+1)} & \text{if } j \leq r \\ 2^{-k} & \text{otherwise} \end{cases}, \quad (3)$$

where $h = kd + r$ is the decomposition with $r \in \{0, \dots, d-1\}$. Indeed, we then have

$$\text{diam}(\mathcal{X}^{(h,i)}) = \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^{(h,i)}} \|\mathbf{x} - \mathbf{x}'\|_2 \leq \sqrt{\sum_{j=1}^d (\delta_j^{(h,i)})^2}.$$

But by (3), for r coordinates $j \in \{1, \dots, r\}$ among the d coordinates $\delta_j^{(h,i)}$ equals $2^{-(k+1)}$ while the $d-r$ remaining coordinates $j \in \{r+1, \dots, d\}$ satisfy $\delta_j^{(h,i)} = 2^{-k}$. Thus, by routine calculations

$$\begin{aligned} \text{diam}(\mathcal{X}^{(h,i)}) &\leq \sqrt{r (2^{-(k+1)})^2 + (d-r) (2^{-k})^2} \\ &= 2^{-k} \sqrt{\frac{r}{4} + d - r} \\ &= \sqrt{d} 2^{-k} \sqrt{1 - \frac{3r}{4d}} \\ &= \sqrt{d} \left(2^{1/d}\right)^{-(dk+r)} 2^{r/d} \sqrt{1 - \frac{3r}{4d}} \end{aligned}$$

But,

$$2^{r/d} \sqrt{1 - \frac{3r}{4d}} \leq \max_{0 \leq u \leq 1} \left\{ 2^u \sqrt{1 - \frac{3u}{4}} \right\} \approx 1.12 \leq \sqrt{2}.$$

The proof is concluded by substituting in the previous bound.

Now, we prove (3) by induction on the depth h . This is true for $h = 0$ as the bin of the root node $\mathcal{X}^{(0,1)}$ equals $[0, 1]^d$ by definition. Besides, let $h \geq 0$ and $i \in \{1, \dots, 2^h\}$. We compute the decomposition $h = kd + r$ with $r \in \{0, \dots, d-1\}$.

We have by step 5.4 of Algorithm 2 that the range of each coordinate $j \neq r + 1$ of the bin of the child node $(h + 1, 2i)$ remains the same

$$\delta_j^{(h+1, 2i)} = \delta_j^{(h, i)} = \begin{cases} 2^{-(k+1)} & \text{if } j \leq r \\ 2^{-k} & \text{if } j \geq r + 2 \end{cases}, \quad (4)$$

and the range of coordinate $r + 1$ is divided by 2,

$$\delta_{r+1}^{(h+1, 2i)} = \delta_{r+1}^{(h, i)} / 2 = 2^{-(k+1)}. \quad (5)$$

Equations (4) and (5) are also true for the second child $(h + 1, 2i - 1)$, and this concludes the induction.

B Lemma 8 and its proof

Lemma 8. *Let $N \geq 1$ be an odd integer. Let \mathcal{T} be a binary tree with N nodes. Then,*

- *its number of inner-nodes equals $N^{\text{in}} = (N - 1)/2$.*
- *the average depth (i.e., distance to the root) of its inner nodes is lower-bounded as*

$$\frac{1}{N^{\text{in}}} \sum_{h=0}^{\infty} h \# \{\text{inner nodes in } \mathcal{T} \text{ of depth } h\} \geq \log_2 \left(\frac{N - 1}{8} \right).$$

Proof. First statement. We proceed by induction. If $N = 1$, there is only one binary tree with one node, the lone leaf, so that $N^{\text{in}} = 0$. Now, if \mathcal{T} is a binary tree with $N \geq 3$ nodes, select an inner node n which is parent of two leaf nodes. Then, replaces the subtree rooted at n by a leaf node. The resulting subtree \mathcal{T}' of \mathcal{T} has $N - 2$ nodes, so that by induction hypothesis \mathcal{T}' has $(N - 3)/2$ inner nodes. But, \mathcal{T}' has also $N^{\text{in}} - 1$ inner nodes. Therefore $N^{\text{in}} = (N - 1)/2$.

Second statement. We note that the average depth is minimized for the equilibrated binary trees, that are such that

- all depths $h \in \{0, \dots, \lfloor \log_2 N^{\text{in}} \rfloor\}$ have exactly 2^h inner nodes;
- no inner nodes has depth $h > \lfloor \log_2 N^{\text{in}} \rfloor$.

Therefore,

$$\frac{1}{N^{\text{in}}} \sum_{h=0}^{\infty} h \# \{\text{inner nodes in } \mathcal{T} \text{ of depth } h\} \geq \frac{1}{N^{\text{in}}} \sum_{h=0}^{\lfloor \log_2 N^{\text{in}} \rfloor} h 2^h$$

Now, we use that $\sum_{i=0}^{n-1} i 2^i = 2^n(n - 2) + 2$ for all $n \geq 1$, which implies because $\lfloor \log_2 N^{\text{in}} \rfloor \geq \log_2 N^{\text{in}} - 1$ and by substituting in the previous bound,

$$\frac{1}{N^{\text{in}}} \sum_{h=0}^{\infty} h \# \{\text{inner nodes in } \mathcal{T} \text{ of depth } h\} \geq \underbrace{\frac{2^{\log_2 N^{\text{in}}}}{N^{\text{in}}}}_{=1} (\log_2 N^{\text{in}} - 2) + \underbrace{\frac{2}{N^{\text{in}}}}_{\geq 0}.$$

This concludes the proof by substituting $N^{\text{in}} = (N - 1)/2$. \square

C Proof of Lemma 6

The proof follows from a simple adaptation of the proof of the regret bound of the exponentially weighted average forecaster—see for instance [CBL06]. By convexity of ℓ and by Hoeffding's inequality, we have at each time step t

$$\ell(\hat{y}_t, y_t) \leq \sum_{d=1}^{D_t} \hat{p}_{d,t} \ell(f_{d,t}, y_t) \leq -\frac{1}{\eta_t} \log \sum_{d=1}^{D_t} \hat{p}_{d,t} e^{-\eta_t \ell(f_{d,t}, y_t)} + \frac{\eta_t}{8}$$

By Jensen's inequality, since $\eta_{t+1} \leq \eta_t$ and thus $x \mapsto x^{\eta_t/\eta_{t+1}}$ is convex

$$\begin{aligned} \frac{1}{D_t} \sum_{d=1}^{D_t} \hat{p}_{d,t} e^{-\eta_t \ell(f_{d,t}, y_t)} &= \frac{1}{D_t} \sum_{d=1}^{D_t} \left(\hat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1} \ell(f_{d,t}, y_t)} \right)^{\frac{\eta_t}{\eta_{t+1}}} \\ &\geq \left(\frac{1}{D_t} \sum_{d=1}^{D_t} \hat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1} \ell(f_{d,t}, y_t)} \right)^{\frac{\eta_t}{\eta_{t+1}}} \end{aligned}$$

Substituting in Hoeffding's bound we get

$$\ell(\hat{y}_t, y_t) \leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \log D_t - \frac{1}{\eta_{t+1}} \log \left(\sum_{d=1}^{D_t} \hat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1} \ell(f_{d,t}, y_t)} \right) + \frac{\eta_t}{8}$$

Now, by definition of the loss update in step 3 of Algorithm 4, for all $d = 1, \dots, D_t$

$$\sum_{k=1}^{D_t} \hat{p}_{k,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1} \ell(f_{k,t}, y_t)} = \frac{D_t}{D_{t+1}} \frac{\hat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1} \ell(f_{d,t}, y_t)}}{\hat{p}_{d,t+1}}$$

which after substitution in the previous bound leads to the inequality

$$\ell(\hat{y}_t, y_t) \leq \ell(f_{d,t}, y_t) + \frac{1}{\eta_{t+1}} \log(D_{t+1} \hat{p}_{d,t+1}) - \frac{1}{\eta_t} \log(D_t \hat{p}_{d,t}) + \frac{\eta_t}{8}.$$

By summing over $t = t_d, \dots, T$, the sum telescopes; using that $\hat{p}_{d,t_d} = 1/D_{t_d}$ by step 3.1.

$$\sum_{t=t_d}^T \ell(\hat{y}_t, y_t) \leq \sum_{t=t_d}^T \ell(f_{d,t}, y_t) + \frac{1}{\eta_{T+1}} \log(D_{T+1} \underbrace{\hat{p}_{d,T+1}}_{\leq 1}) - \frac{1}{\eta_{t_d}} \log(D_{t_d} \underbrace{\hat{p}_{d,t_d}}_{=1}) + \frac{1}{8} \sum_{t=t_d}^T \eta_t,$$

which concludes the proof of the first statement. The second statement of the theorem is because

$$\frac{1}{2} \sum_{t=1}^T \eta_t = \sum_{t=1}^T \frac{1}{\sqrt{t}} = 1 + \sum_{t=2}^T \frac{1}{\sqrt{t}} \leq 1 + \int_1^T \frac{1}{\sqrt{t}} dt \leq 2\sqrt{T}.$$

D Proof of Lemma 7

The proof is performed in two steps.

Step 1: Lipschitz \rightarrow Continuous. First, the Stone-Weierstrass theorem entails that any continuous function $f : \mathcal{X} \rightarrow \mathbb{R}$ from a compact metric space \mathcal{X} to \mathbb{R} is the uniform limit of Lipschitz functions, see e.g., [Geo67]. Thus, the dominated convergence theorem yields

$$\inf_{f \in \mathcal{L}} \mathbb{E}[\ell(f(X), Y)] = \inf_{f \in \mathcal{C}} \mathbb{E}[\ell(f(X), Y)],$$

where \mathcal{L} denotes the set of Lipschitz functions from \mathcal{X} to \mathbb{R} and \mathcal{C} is the set of continuous functions from \mathcal{X} to \mathbb{R} .

Step 2: Continuous \rightarrow Borel. Second, by the version of Lusin's theorem stated in Theorem 4, we can approximate any measurable function by continuous functions (this is where regularity is used).

Let $\delta, \varepsilon > 0$ and $f : \mathcal{X} \rightarrow [0, 1]$ be a Borel function. By Theorem 4, there exists a continuous function $g : \mathcal{X} \rightarrow [0, 1]$ such that

$$\mathbb{P}_X\{|f - g| \geq \delta\} \leq \varepsilon.$$

Then by Jensen's inequality, and since

$$\begin{aligned} \Delta &\triangleq \left| \mathbb{E}[\ell(f(X), Y)] - \mathbb{E}[\ell(g(X), Y)] \right| \leq \mathbb{E}\left[\left| \ell(f(X), Y) - \ell(g(X), Y) \right| \right] \\ &\leq \underbrace{\mathbb{P}_X\{|f - g| \geq \delta\}}_{\leq \varepsilon} + \underbrace{\mathbb{E}\left[M|f(X) - g(X)| \mathbf{1}_{\{|f(X) - g(X)| \leq \delta\}} \right]}_{\leq M\delta}, \end{aligned}$$

where the second inequality is because ℓ takes values in $[0, 1]$ and is M -Lipschitz in its first argument. Thus $\Delta \leq \varepsilon + M\delta$, which concludes the proof since this is true for arbitrary small values of ε and δ .

Theorem 4 (Lusin). *If \mathcal{X} is a convex and compact subset of a normed space, equipped with a regular probability measure μ , then for every measurable function $f : \mathcal{X} \rightarrow [0, 1]$ and for every $\delta, \varepsilon > 0$, there exists a continuous function $g : \mathcal{X} \rightarrow [0, 1]$ such that*

$$\mu\{|f - g| \geq \delta\} \leq \varepsilon.$$

The proof of Theorem 4 can be easily derived from the proof of [SL07, Proposition 25].

E Proof of Theorem 3

In this proof, apart from the use of Breiman's generalized ergodic theorem in the beginning and the martingale convergence theorem in the end (as exhibited in [GLF01,GO07,BBGO10,BP11]), we resort to new arguments.

Let $d \geq 1$ and $L \geq 0$. Then, by assumption and by exchanging lim sup and inf,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \left(\sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \leq \inf_{f \in \mathcal{L}_L^d} \limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(f(Y_{t-d}^{t-1}), Y_t) \right).$$

Because ℓ is bounded over $[0, 1]^2$ and thus integrable, Breiman's generalized ergodic theorem (see [Bre57]) entails that the right-term converges: almost surely,

$$\lim_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(f(Y_{t-d}^{t-1}), Y_t) \right) = \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)]$$

and thus,

$$\limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \leq \inf_{f \in \mathcal{L}_L^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)].$$

By letting $L \rightarrow \infty$ in the inequality above, we get

$$\limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) \leq \inf_{f \in \mathcal{L}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)].$$

By Lemma 7 the infimum over all continuous functions equals the infimum over the set \mathcal{B}^d of Borel functions. Therefore,

$$\begin{aligned} \limsup_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) \right) &\leq \inf_{f \in \mathcal{B}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0)] \\ &\leq \mathbb{E} \left[\underbrace{\inf_{f \in \mathcal{B}^d} \mathbb{E}[\ell(f(Y_{-d}^{-1}), Y_0) | Y_{-d}^{-1}]}_{\triangleq Z_d} \right], \end{aligned}$$

where the second inequality is by the measurable selection theorem—see Theorem 8 in Appendix I of [Alg94]. Now, we remark that (Z_d) is a bounded supermartingale with respect to the family of sigma algebras $(\sigma(Y_{-d}^{-1}))_{d \geq 1}$. Indeed, the function $\inf_{f \in \mathcal{B}^{d+1}}(\cdot)$ is concave, thus conditional Jensen's inequality

$$\begin{aligned} \mathbb{E}[Z_{d+1} | Y_{-d}^{-1}] &\leq \inf_{f \in \mathcal{B}^{d+1}} \mathbb{E} \left[\mathbb{E}[\ell(f(Y_{-(d+1)}^{-1}), Y_0) | Y_{-(d+1)}^{-1}] | Y_{-d}^{-1} \right] \\ &= \inf_{f \in \mathcal{B}^{d+1}} \mathbb{E}[\ell(f(Y_{-(d+1)}^{-1}), Y_0) | Y_{-d}^{-1}] \end{aligned}$$

Now, we note that

$$\inf_{f \in \mathcal{B}^{d+1}} \mathbb{E} \left[\ell \left(f(Y_{-(d+1)}^{-1}), Y_0 \right) \middle| Y_{-d}^{-1} \right] \leq \inf_{f' \in \mathcal{B}^d} \mathbb{E} \left[\ell \left(f'(Y_{-d}^{-1}), Y_0 \right) \middle| Y_{-d}^{-1} \right] = Z_d,$$

which yields $\mathbb{E}[Z_{d+1} | Y_{-d}^{-1}] \leq Z_d$. Thus, the martingale convergence theorem (see e.g. [Cho65]) implies that Z_d converges almost surely and in \mathbb{L}_1 . Thus,

$$\lim_{d \rightarrow \infty} \mathbb{E}[Z_d] = \mathbb{E} \left[\inf_{f \in \mathcal{B}^\infty} \mathbb{E} \left[\ell(f(Y_{-\infty}^{-1}), Y_0) \middle| Y_{-\infty}^{-1} \right] \right] = L^*,$$

which yields the stated result $\limsup_T \sum_{t=1}^T \ell(\hat{Y}_t, Y_t) / T = L^*$.